

APPLICATION
FOR
UNITED STATES LETTERS PATENT

DEPARTMENT OF COMMERCE
U.S. PATENT AND TRADEMARK OFFICE

TITLE: MACHINE TRANSLATION TECHNIQUES
APPLICANT: DANIEL MARCU

CERTIFICATE OF MAILING BY EXPRESS MAIL

Express Mail Label No. EL688319725US

I hereby certify under 37 CFR §1.10 that this correspondence is being deposited with the United States Postal Service as Express Mail Post Office to Addressee with sufficient postage on the date indicated below and is addressed to the Commissioner for Patents, Washington, D.C. 20231.

Date of Deposit

May 11, 2001

Signature



Gildardo Vargas
Typed or Printed Name of Person Signing Certificate

TITLE

MACHINE TRANSLATION TECHNIQUES

Related Application

5 This application claims the benefit of, and incorporates
herein, U.S. Provisional Patent Application No. 60/203,643, filed
May 11, 2000.

Origin of Invention

10 The research and development described in this application
were supported by the NSA under grant number MDA904-97-0262 and
by DARPA / ITO under grant number MDA904-99-C-2535. The US
government may have certain rights in the claimed inventions.

Field of the Invention

15 The present application relates to computational linguistics
and more particularly to machine translation techniques. More
specifically, the present application describes techniques for
performing decoding of a source text segment into a target text
20 segment and for rewriting trees from a first linguistic space
 into another linguistic space.

Background and Summary

Machine translation (MT) is the automatic translation, for example, using a computer system, from a first language (e.g., French) into another language (e.g., English). Systems that 5 perform MT techniques are said to "decode" the source language into the target language. From the end-user's perspective, the MT process is relatively straight-forward. As shown in Fig. 1A, the MT 102 receives as input a source sentence 100, for example, in French (e.g., "ce ne est pas juste"), and after processing 10 the input sentence, outputs the equivalent decoded sentence in the target language - in this example, English ("it is not fair").

One type of conventional MT decoder is the "stack decoder" such as described in U.S. Patent No. 5,477,451 (Brown et al.), 15 entitled "Method and System for Natural Language Translation."

In a stack decoder, the universe of possible translations are organized into a graph structure and then exhaustively searched until an optimal solution (translation) is found. Although stack decoders tend to produce good results, they do so at a 20 significant cost - namely, maintaining and searching a large potential solution space such as used by stack decoders is expensive, both computationally and space-wise (e.g., in terms of computer memory). Accordingly, the present inventor recognized that an iterative, incremental decoding technique could produce

optimal, or near optimal, results while considerably reducing the computational and space requirements. This decoder is referred to herein as a "greedy" decoder or, equivalently, as a "fast decoder." The term "greedy" refers to techniques that produce 5 solutions based on myopic optimization - that is, given a partial solution, produce as a next estimate a new solution that improves the objective the most. Put another way, a greedy algorithm typically starts out with an approximate solution and then tries to improve it incrementally until a satisfactory solution is 10 reached.

Implementations of the greedy decoder may include various combinations of the following features.

In one aspect, machine translation (MT) decoding involves receiving as input a text segment (e.g., a clause, a sentence, a 15 paragraph or a treatise) in a source language to be translated into a target language, generating an initial translation (e.g., either a word-for-word or phrase-for-phrase gloss) as a current target language translation, applying one or more modification operators to the current target language translation to generate 20 one or more modified target language translations, determining whether one or more of the modified target language translations represents an improved translation in comparison with the current target language translation, setting a modified target language translation as the current target language translation, and

repeating these steps until occurrence of a termination condition.

Applying one or more modification operators may involve changing the translation of one or two words in the current target language translation the translation. Alternatively, or in addition, applying one or more modification operators may include (i) changing a translation of a word in the current target language translation and concurrently (ii) inserting another word at a position that yields an alignment of highest probability between the source language text segment and the current target language translation. The inserted other word may have a high probability of having a zero-value fertility.

Applying one or more modification operators may include deleting from the current target language translation a word having a zero-value fertility; and/or modifying an alignment between the source language text segment and the current target language translation by swapping non-overlapping target language word segments in the current target language translation; and/or modifying an alignment between the source language text segment and the current target language translation by (i) eliminating a target language word from the current target language translation and (ii) linking words in the source language text segment.

In various embodiments, applying modification operators may include applying two or more of the following: (i) changing the

translation of one or two words in the current target language translation; (ii) changing a translation of a word in the current target language translation and concurrently inserting another word at a position that yields an alignment of highest probability between the source language text segment and the current target language translation, the inserted other word having a high probability of having a zero-value fertility; (iii) deleting from the current target language translation a word having a zero-value fertility; (iv) modifying an alignment between the source language text segment and the current target language translation by swapping non-overlapping target language word segments in the current target language translation; and/or (v) modifying an alignment between the source language text segment and the current target language translation by eliminating a target language word from the current target language translation and linking words in the source language text segment.

Determining whether one or more of the modified target language translations represents an improved translation in comparison with the current target language translation may include calculating a probability of correctness for each of the modified target language translations.

The termination condition may include a determination that a probability of correctness of a modified target language

translation is no greater than a probability of correctness of the current target language translation. The termination condition may be the occurrence of a completion of a predetermined number of iterations and/or the lapse of a 5 predetermined amount of time.

In another aspect, a computer-implemented machine translation decoding method may, for example, implement a greedy decoding algorithm that iteratively modifies a target language translation of a source language text segment (e.g., a clause, a sentence, a paragraph, or a treatise) until an occurrence of a 10 termination condition (e.g., completion of a predetermined number of iterations, lapse of a predetermined period of time, and/or a determination that a probability of correctness of a modified translation is no greater than a probability of correctness of a 15 previous translation.)

The MT decoding method may start with an approximate target language translation and iteratively improve the translation with each successive iteration. The approximate target language translation may be, for example, a word-for-word or phrase-for- 20 phrase gloss, or the approximate target language translation may be a predetermined translation selected from among a plurality of predetermined translations.

Iteratively modifying the translation may include incrementally improving the translation with each iteration, for

example, by applying one or more modification operations on the translation.

The one or more modification operations comprises one or more of the following operations: (i) changing one or two words 5 in the translation; (ii) changing a translation of a word and concurrently inserting another word at a position that yields an alignment of highest probability between the source language text segment and the translation, the inserted other word having a high probability of having a zero-value fertility; (iii) deleting 10 from the translation a word having a zero-value fertility; (iv) modifying an alignment between the source language text segment and the translation by swapping non-overlapping target language word segments in the translation; and (v) modifying an alignment between the source language text segment and the translation by 15 eliminating a target language word from the translation and linking words in the source language text segment.

In another aspect, a machine translation decoder may include a decoding engine comprising one or more modification operators to be applied to a current target language translation to 20 generate one or more modified target language translations; and a process loop to iteratively modify the current target language translation using the one or more modification operators. The process loop may terminate upon occurrence of a termination condition. The process loop may control the decoding engine to

incrementally improve the current target language translation with each iteration.

The MT decoder may further include a module (including, for example, a language model and a translation model) for

5 determining a probability of correctness for a translation.

The process loop may terminate upon a determination that a probability of correctness of a modified translation is no greater than a probability of correctness of a previous translation, and/or upon completion of a predetermined number of iterations; and/or after lapse of a predetermined period of time.

One or more of the following advantages may be provided by the greedy decoder as described herein. The techniques and methods described here may result in a MT decoder that performs with high accuracy, high speed and relatively low computational

15 and space costs. The greedy decoder can be modified as desired to perform a full set of sentence modification operations or any subset thereof. This gives a system designer and/or end-user considerable flexibility to tailor the decoder's speed, accuracy and/or other performance characteristics to match desired

20 objectives or constraints. The use of a set of basic modification operations, each able to be used as a standalone operator or in conjunction with the others, further enhances this flexibility. Moreover, the use of independent standalone operators as constituents of the decoding engine makes the

decoder extensible and scalable. That is, different or additional modification operators can be used to suit the objectives or constraints of the system designer and/or end-user.

In conjunction with MT research and related areas in
5 computational linguistics, researchers have developed and frequently use various types of tree structures to graphically represent the structure of a text segment (e.g., clause, sentence, paragraph or entire treatise). Two basic tree types include (1) the syntactic tree, which can be used to graphically represent the syntactic relations among components of a text segment, and (2) the rhetorical tree (equivalently, the rhetorical structure tree (RST) or the discourse tree), which can be used to graph the rhetorical relationships among components of a text segment. Rhetorical structure trees (also referred to as
10 discourse trees) are discussed in detail in William C. Mann and Sandra A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," Text, 8(3):243-281
15 (1988).

The example shown in Fig. 6 illustrates the types of
20 structures that may be present in a rhetorical structure tree for a text fragment. The leaves of the tree correspond to elementary discourse units ("edus") and the internal nodes correspond to contiguous text spans. Each node in a rhetorical structure tree is characterized by a "status" (i.e., either "nucleus" or

"satellite") and a "rhetorical relation," which is a relation that holds between two non-overlapping text spans. In Fig. 6, nuclei are represented by straight lines while satellites are represented by arcs.

5 The present inventor recognized that significant differences exist between the rhetorical structures of translations of a text in different languages (e.g., Japanese and English).

Accordingly, to improve MT quality, and as a component of a
10 larger MT system, the present inventor developed techniques for
automatically rewriting (e.g., using a computer system)
rhetorical structures from one language into another, for
example, rewriting a rhetorical tree for a text segment in
Japanese into a rhetorical tree for a counterpart text segment in
English.

15 Implementations of the disclosed tree rewriting techniques
may include various combinations of the following features.

In one aspect, automatically generating a tree (e.g., either
a syntactic tree or a discourse tree) involves receiving as input
a tree corresponding to a source language text segment, and
20 applying one or more decision rules to the received input to
generate a tree corresponding to a target language text segment.

In another aspect, a computer-implemented tree generation
method may include receiving as input a tree corresponding to a
source language text segment (e.g., clause, sentence, paragraph,

or treatise), and applying one or more decision rules (e.g., a sequence of decision rules that collectively represent a transfer function) to the received input to generate a tree corresponding to a target language text segment, which potentially may be a 5 different type of text segment.

The tree generation method further may include automatically determining the one or more decision rules based on a training set, for example, a plurality of input-output tree pairs and a mapping between each of the input-output tree pairs. The mapping between each of the input-output tree pairs may be a mapping between leaves of the input tree and leaves of the paired output tree. Mappings between leaves of input-output tree pairs can be one-to-one, one-to-many, many-to-one, or many-to-many.

10 Automatically determining the one or more decision rules may include determining a sequence of operations that generates an output tree when applied to the paired input tree. Determining a sequence of operations may include using a plurality of predefined operations that collectively are sufficient to render any input tree into the input tree's paired output tree. The 15 plurality of predefined operations comprise one or more of the following: a shift operation that transfers an elementary discourse tree (edt) from an input list into a stack; a reduce operation that pops two edts from a top of the stack, combines the two popped edts into a new tree, and pushes the new tree on 20

the top of the stack; a break operation that breaks an edt into a predetermined number of units; a create-next operation that creates a target language discourse constituent that has no correspondent in the source language tree; a fuse operation that 5 fuses an edt at the top of the stack into the preceding edt; a swap operation that swaps positions of edts in the input list; and an assignType operation that assigns one or more of the following types to edts: Unit, MultiUnit, Sentence, Paragraph, MultiParagraph, and Text.

10 The plurality of predefined operations may represent a closed set that includes the shift operation, the reduce operation, the break operation, the create-next operation, the fuse operation, the swap operation and the assignType operation.

15 Determining a sequence of operations may result in a plurality of learning cases, one learning case for each input-output tree pair. In that case, the tree generation method may further include associating one or more features with each of the plurality of learning cases based on context. The associated features may include one or more of the following: operational 20 and discourse features, correspondence-based features, and lexical features.

The tree generation method may further include applying a learning program (e.g., C4.5) to the plurality of learning cases to generate the one or more decision rules

In another aspect, a computer-implemented tree generation module may include a predetermined set of decision rules that, when applied to a tree (e.g., syntactic or discourse) corresponding to a source language text segment, generate a tree corresponding to a target language text segment. The predetermined set of decision rules may define a transfer function between source language trees and target language trees.

In another aspect, determining a transfer function between trees (e.g., syntactic or discourse) of different types may include generating a training set comprising a plurality of tree pairs and a mapping between each tree pair, each tree pair comprises a source tree and a corresponding target tree, and generating a plurality of learning cases by determining, for each tree pair, a sequence of operations that result in the target tree when applied to the source tree; and generating a plurality of decision rules by applying a learning algorithm to the plurality of learning cases.

Determining a transfer function between trees of different types further may include, prior to generating the plurality of decision rules, associating one or more features with each of the learning cases based on context.

In another aspect, a computer-implemented discourse-based machine translation system may include a discourse parser that parses the discourse structure of a source language text segment

and generates a source language discourse tree for the text segment; a discourse-structure transfer module that accepts the source language discourse tree as input and generates as output a target language discourse tree; and a mapping module that
5 maps the target language discourse tree into a target text segment. The discourse-structure transfer module may include a plurality of decision rules generated from a training set of source language-target language tree pairs.

One or more of the following advantages may be provided by
10 tree rewriting as described herein. The techniques and methods described here may result in a tree rewriting capability that allows users (e.g., human end-users such as linguistic researchers or computer processes such as MT systems) to automatically have a tree for a text segment in a source language
15 rewritten, or translated, into a tree for the text segment translated into a target language. This functionality is useful both in its standalone form and as a component of a larger system, such as in a discourse-based machine translation system. Moreover, because the tree rewriter described here automatically
20 learns how to rewrite trees from one language into another, the system is easy and convenient to use.

The mapping scheme used in training the tree rewriter also provides several advantages. For example, by allowing any arbitrary groupings (e.g., one-to-one, one-to-many, many-to-one,

many-to-many) between leaves in the source and target trees, the flexibility, richness and robustness of the resulting mappings are enhanced.

5 The enhanced shift-reduce operations used in training the tree rewriter also provide several advantages. For example, the set of basic operations that collectively are sufficient to render any input tree into its paired output tree provides a powerful yet compact tool for rewriting tree structures.

10 The details of one or more embodiments are set forth in the accompanying drawings and the description below. Other features, objects, and advantages of the invention will be apparent from the description and drawings, and from the claims.

Drawing Descriptions

15 These and other aspects of the invention will now be described in detail with reference to the accompanying drawings, wherein:

Figure 1A shows a block diagram of machine translation from a user's perspective.

20 Figure 1B shows an example of word-level alignment.

Figure 2 shows a flowchart of the operation of an embodiment of the greedy decoder.

Figure 3 shows an example of the greedy decoder producing an English translation of a French sentence.

Figure 4 shows an example of output that a user sees as the greedy decoder produces an English translation of a French sentence.

5 Figure 5 is a table showing a comparison between different decoders using a trigram language model.

Figure 6 shows an example of a rhetorical structure tree.

Figure 7 is an example of a Japanese source sentence.

Figure 8 is the discourse structure of the Japanese source sentence in Fig. 7.

10 Figure 9 is the discourse structure of an English target sentence translated from the Japanese source sentence of Fig. 11.

Figure 10 shows a block diagram of the tree rewriter.

Figure 11 shows a block diagram of how a tree rewriter can be used as a subsystem of a larger system.

15 Figure 12 shows a block diagram of a discourse-based machine translation system with the tree rewriter as a subsystem.

Figure 13 is a flowchart of a procedure for building a tree rewriter.

20 Figure 14 shows an example of incremental tree reconstruction.

Figure 15 is a graph of the learning curve for the Relation-Reduce classifier.

Detailed Description

Greedy Decoder

A statistical MT system that translates, for example, French sentences into English, can be divided into three parts: (1) a 5 language model (LM) that assigns a probability $P(e)$ to any English string, (2) a translation model (TM) that assigns a probability $P(f|e)$ to any pair of English and French strings, and (3) a decoder. The decoder takes a previously unseen sentence f and tries to find the e that maximizes $P(e|f)$, or equivalently 10 maximizes $P(e) \cdot P(f|e)$.

Brown et al., "The mathematics of statistical machine translation: Parameter estimation", *Computational Linguistics*, 19(2), 1993, introduced a series of TMs based on word-for-word substitution and re-ordering, but did not include a decoding 15 algorithm. If the source and target languages are constrained to have the same word order (by choice or through suitable pre-processing), then the linear Viterbi algorithm can be applied as described in Tillmann et al., "A DP-based search using monotone alignments in statistical translation", In *Proc. ACL*, 1997. If 20 re-ordering is limited to rotations around nodes in a binary tree, then optimal decoding can be carried out by a high-polynomial algorithm (Wu, "A polynomial-time algorithm for statistical machine translation", In *Proc. ACL*, 1996). For arbitrary word-reordering, the decoding problem is NP-complete

(nondeterministic polynomial time complete) (Knight, "Decoding complexity in word-replacement translation models", *Computational Linguistics*, 25(4), 1999).

One strategy (Brown et al., "Method and system for natural language translation," U.S. Patent 5,477,451, 1995; Wang et al., "Decoding algorithm in statistical machine translation", In *Proc. ACL, 1997*) is to examine a large subset of likely decodings and chose just from that. Of course, it is possible to miss a good translation this way.

Thus, while decoding is a clear-cut optimization task in which every problem instance has a right answer, it is hard to come up with good answers quickly. The following sets forth details of a fast greedy decoder and compares its performance with that of a traditional stack decoder.

In developing the greedy decoder IBM Model 4 was used, which revolves around the notion of a word alignment over a pair of sentences (see Figure 1B). A word alignment assigns a single home (English string position) to each French word. If two French words align to the same English word, then that English word is said to have a fertility of two. Likewise, if an English word remains unaligned-to, then it has fertility zero. The word alignment in Figure 1B is shorthand for a hypothetical stochastic process by which an English string gets converted into a French string. There are several sets of decisions to be made.

First, every English word is assigned a fertility. These assignments are made stochastically according to a table $n(\emptyset|e_i)$. Any word with fertility zero is deleted from the string, any word with fertility two is duplicated, etc. If a word has fertility 5 greater than one, it is called *very fertile*.

After each English word in the new string, the fertility of an invisible English NULL element with probability p_1 (typically about 0.02) is incremented. The NULL element ultimately produces "spurious" French words.

10 Next, a word-for-word replacement of English words (including NULL) by French words is performed, according to the table $t(f_j|e_i)$.

Finally, the French words are permuted. In permuting, IBM Model 4 distinguishes between French words that are *heads* (the 15 leftmost French word generated from a particular English word), *non-heads* (non-leftmost, generated only by very fertile English words), and *NULL-generated*.

Heads. The head of one English word is assigned a French 20 string position based on the position assigned to the previous English word. If an English word E_{i-1} translates into something at French position j , then the French head word of e_i is stochastically placed in French position k with distortion probability $d_i(k-j|class(e_{i-1}))$, $class(f_k)$, where "class" refers

to automatically determined word classes for French and English vocabulary items. This relative offset $k-j$ encourages adjacent English words to translate into adjacent French words. If e_{i-1} is infertile, then j is taken from e_{i-2} , etc. If e_{i-1} is very 5 fertile, then j is the average of the positions of its French translations.

10 **Non-heads.** If the head of English word e_i is placed in French position j , then its first non-head is placed in French position $k (>j)$ according to another table $d_{i,1}(k-j|class (f_k))$. The next non-head is placed at position q with probability $d_{i,1}(q-k|class (f_q))$, and so forth.

15 **NULL-generated.** After heads and non-heads are placed, NULL-generated words are permuted into the remaining vacant slots randomly. If there are \emptyset_0 NULL-generated words, then any placement scheme is chosen with probability $1/\emptyset_0!$.

20 These stochastic decisions, starting with e , result in different choices of f and an alignment of f with e . e is mapped onto a particular $\langle a, f \rangle$ pair with probability:

$$P(a, f|e) =$$

$$\begin{aligned}
 & \prod_{i=1}^J n(\phi_i | e_i) x \prod_{i=1}^J \prod_{k=1}^{\phi_i} t(\tau_{ik} | e_i) x \\
 & \prod_{i=1, \rho_i > 0}^J d_i(\pi_{ii} - c_{\rho_i} | \text{class}(e_{\rho_i}), \text{class}(\tau_{ii})) \\
 & \prod_{i=1}^J \prod_{k=2}^{\phi_i} d_{i-1}(\pi_{ik} - \pi_{i(k-1)} | \text{class}(\tau_{ik})) x \\
 & \binom{m - \phi_0}{\phi_0} p_1^{\phi_0} (1 - p_1)^{m - 2\phi_0} x \\
 & \prod_{k=1}^{\phi_0} t(\tau_{0k} | \text{NULL})
 \end{aligned}$$

where the factors separated by x symbols denote fertility, translation, head permutation, non-head permutation, null-fertility, and null-translation probabilities. The symbols in this formula are: J (the length of e), m (the length of f), e_i (the i^{th} English word in e), E_0 (the NULL word), ϕ_i (the fertility of e_{ii}), ϕ_0 (the fertility of the NULL word), τ_{ik} (the k^{th} French word produced by e_i in a), π_{ik} (the position of τ_{ik} in f), ρ_i (the position of the first fertile word to the left of e_i in a), c_{ρ_i}

10 (the ceiling of the average of all $\pi_{\rho_i k}$ for ρ_i , or 0 if ρ_i is undefined).

15 In view of the foregoing, given a new sentence f, then an optimal decoder will search for an e that maximizes $P(e|f) \approx P(e) \cdot P(f|e)$. Here, $P(f|e)$ is the sum of $P(a, f|e)$ over all possible alignments a. Because this sum involves significant computation,

typically it is avoided by instead searching for an $\langle e, a \rangle$ pair that maximizes $P(e, a|f) \approx P(e) \cdot P(a, f|e)$. It is assumed that the language model $P(e)$ is a smoothed n-gram model of English.

Figure 2 is a flowchart of the operation of an embodiment of 5 a greedy decoder for performing MT. As shown therein, the first step 200 is to receive an input sentence to be translated.

Although in this example, the text segment being translated is a sentence, virtually any other text segment could be used, for example, clauses, paragraphs or entire treatises.

10 In step 202, as a first approximation of the translation, the greedy decoder generates a "gloss" of the input sentence, essentially a word-for-word translation. The gloss is constructed by aligning each French word f_j with its most likely English translation e_{f_j} ($e_{f_j} = \text{argmax}_e t(e | f_j)$). For example, 15 in translating the French sentence "Bien entendu, il parle de une belle victoire", the greedy decoder initially assumes that a good translation is ``Well heard, it talking a beautiful victory'' because the best translation of "bien" is "well", the best translation of "entendu" is "heard", and so on. The 20 alignment corresponding to this translation is shown at the top of Figure 3.

In step 204, the decoder estimates the probability of correctness of the current translation, $P(c)$.

After the initial alignment is generated in step 202, the greedy decoder tries to improve it in step 206. That is, the decoder tries to find an alignment (and implicitly, a translation) of higher probability by applying one or more 5 sentence modification operators, described below. The use of a word-level alignment and the particular operators described below were chosen for this particular embodiment. However, alternative embodiments using different statistical models may benefit from different or additional operations.

10 The following operators collectively make-up the decoder's translation engine, and include the following:

translateOneOrTwoWords (j_1, e_1, j_2, e_2)

15 This operation changes the translation of one or two French words, those located at positions j_1 and j_2 , from ef_{j_1} and ef_{j_2} into e_1 and e_2 . If ef_{j_1} is a word of fertility 1 and e_k is NULL, then ef_{j_1} is deleted from the translation. If ef_{j_1} is the NULL word, the word e_k is inserted into the translation at the position that yields an alignment of highest probability. If $ef_{j_1} = e_1$ or $ef_{j_2} =$ 20 e_2 , then this operation amounts to changing the translation of a single word.

translateAndInsert (j, e₁, e₂)

This operation changes the translation of the French word located at position j from e_{fj} into e_1 and simultaneously inserts word e_2 at the position that yields the alignment of highest probability. Word e_2 is selected from an automatically derived list of 1024 words with high probability of having fertility 0. When $e_{fj} = e_1$, this operation amounts to inserting a word of fertility 0 into the alignment.

removeWordOfFertility0 (i)

This operation deletes the word of fertility 0 at position i in the current alignment.

swapSegments (i₁, i₂, j₁, j₂)

This operation creates a new alignment from the old one by swapping non-overlapping English word segments $[i_1, i_2]$ and $[j_1, j_2]$. During the swap operation, all existing links between English and French words are preserved. The segments can be as small as a word or as long as $|e|-1$ words, where $|e|$ is the length of the English sentence.

joinWords (i_1, i_2)

This operation eliminates from the alignment the English word at position i_1 (or i_2) and links the French words generated by e_{i_1} (or e_{i_2}) to e_{i_2} (or e_{i_1}).

5 At step 208, the decoder estimates the probabilities of correctness, $P(M_1) \dots P(M_n)$, for each of the results of the sentence modification operations. That is, the probability for each new resulting translation is determined

At step 210, the decoder determines whether any of the new translations are better than the current translation by comparing their respective probabilities of correctness. If any of the new translations represents a better solution than the current translation, then the best new translation (that is, the translation solution having the highest probability of correctness) is set as the current translation in step 214 and the decoding process returns to step 206 to perform one or more of the sentence modification operations on the new current translation solution.

Steps 206, 208, 210 and 214 repeat until the sentence 20 modification operations cease (as determined in step 210) to produce translation solutions having higher probabilities of correctness, at which point, the decoding process halts at step 212 and the current translation is output as the final decoding solution. Alternatively, the decoder could cease after a

predetermined number of iterations chosen, for example, either by a human end-user or by an application program using the decoder as a translation engine.

Accordingly, in a stepwise fashion, starting from the

5 initial gloss, the greedy decoder uses a process loop (e.g., as shown in Fig. 2, steps 206, 208, 210 and 214) to iterate exhaustively over all alignments that are one operation away from the alignment under consideration. At every step, the decoder chooses the alignment of highest probability, until the
10 probability of the current alignment can no longer be improved. When it starts from the gloss of the French sentence "Bien entendu, il parle de une belle victoire.", for example, the greedy decoder alters the initial alignment incrementally as shown in Figure 3, eventually producing the translation "Quite
15 naturally, he talks about a great victory.". In this process, the decoder explores a total of 77421 distinct alignments / translations, of which "Quite naturally, he talks about a great victory." has the highest probability.

In step 206 of the decoding process, either all of the five
20 sentence modification operations can be used or any subset thereof may be used to the exclusion of the others, depending on the preferences of the system designer and/or end-user. For example, the most time consuming operations in the decoder are ***swapSegments***, ***translateOneOrTwoWords***, and ***translateAndInsert***.

SwapSegments iterates over all possible non-overlapping span pairs that can be built on a sequence of length $|e|$.

TranslateOneOrTwoWords iterates over $|f|^2 |x| |t|^2$ alignments, where $|f|$ is the size of the French sentence and $|t|$ is the 5 number of translations associated with each word (in this implementation, this number is limited to the top 10 translations). **TranslateAndInsert** iterates over $|f| |x| |t| |x| |z|$ alignments, where $|z|$ is the size of the list of words with high probability of having fertility 0 (1024 words in this implementation). Accordingly, the decoder could be designed to omit one or more of these slower operations in order to speed up decoding, but potentially at the cost of accuracy.

Alternatively, or in addition, the decoder could be designed to use different or additional sentence modification operations 15 according to the objectives of the system designer and/or end-user.

An advantage of the greedy decoder comes from its speed. As the experiments described below demonstrate, the greedy decoder can produce translations faster than any other decoder. The 20 greedy decoder is an instance of an "anytime algorithm" - the longer it runs, the better the translation it finds. One potential tradeoff of the greedy decoder relates to the size of the solution space that it explores, which is relatively small.

The farther away a good translation is from the initial gloss, the less likely the greedy decoder is to find it.

Fig. 4 shows another example of the greedy decoder in action in which an acceptable solution is reached in four iterations.

5 As shown therein, the input sentence to be translated is "ce ne
est pas juste." The decoder uses the initial gloss "that not
is not fair." and determines that this translation solution
(Iteration 1) has a probability of correctness ("Aprob" - the
product of LMprob and TMprob) of 1.13162e-22, based on a language
10 model probability (LMprob) of 2.98457e-14 and a translation model
probability (TMprob) of 3.79156e-09.

In the second iteration, the decoder changes the first
instance of the word "not" in the translation to "is" by
applying the **translateOneOrTwoWords** operation, resulting in a new
15 translation solution "that is is not fair", having the
probabilities shown in Fig. 4, Iteration 2. In the third
iteration, the decoder applies the **removeWordOffertility0**
operation and drops one instance of the word "is" in the
translation, resulting in a new translation solution "that is
20 not fair", having the probabilities shown in Fig. 4, Iteration
3. In the fourth and final iteration, the decoder applies the
translateOneOrTwoWords operation again to change the word
"that" in the translation to "it", resulting in the final

translation solution "it is not fair", having the probabilities shown in Fig. 4, Iteration 4.

To determine the performance of the greedy decoder, a series of experiments was performed. In all experiments, decoding was 5 performed using only the top 10 translations of a word, as determined during training, and a list of 1024 words of fertility 0, which was also extracted automatically from the test corpus.

In experiments to determine the accuracy and compare the speed of the greedy decoder described to a conventional stack 10 decoder (such as described in U.S. Patent No. 5,477,451 to Brown et al., a test collection of 505 sentences was used, uniformly distributed across the lengths 6, 8, 10, 15, and 20. Decoders were evaluated with respect to (1) speed, and (2) translation 15 accuracy.

The results in the table shown in Fig. 5, obtained with 20 decoders that use a trigram language model, show that the greedy decoding algorithm is an advantageous alternative to the traditional stack decoding algorithm. Even when the greedy decoder used an optimized-for-speed set of operations (that is, a subset of the total set of five sentence modification operations discussed above) in which at most one word is translated, moved, or inserted at a time - which is labeled "greedy*" in Fig. 5 - the translation accuracy is affected only slightly. In contrast, the translation speed increases with at least one order of 25

magnitude. Depending on the application of interest, one may choose to use a slow decoder that provides optimal results or a fast, greedy decoder that provides non-optimal, but acceptable results.

5 Alternative embodiments for the greedy decoder are possible. For example, the greedy decoder could start with multiple different initial translations (e.g., different variations on the gloss used in step 202 in Fig. 2) and then run the greedy decoding algorithm (i.e., steps 204-214 in Fig. 2) on each of the 10 different initial translations in parallel. For example, the greedy decoder could start with an initial, approximate translation selected from among multiple translated phrases stored in a memory. In the end, the best translation could be selected. This parallel translation of different initial 15 solutions could result in more accurate translations.

Tree Rewriter

Almost all conventional MT systems process text one sentence at a time. Because of this limited focus, MT systems generally 20 cannot re-group and re-order the clauses and sentences of an input text to achieve the most natural rendering in a target language. Yet, even between languages as close as English and French, there is a 10% mismatch in number of sentences - what is said in two sentences in one language is said in only one, or in

three, in another (Gale et al., "A program for aligning sentences in bilingual corpora," *Computational Linguistics*, 19(1):75-102, 1993). For distant language pairs, such as Japanese and English, the differences are more significant.

5 Consider, for example, the Japanese sentence shown in Fig. 7 ("text (1)"). The following ("text(2)") is a word-by-word "gloss" of text (1):

10 (2) [The Ministry of Health and Welfare last year revealed] [population of future estimate according to²] [in future 1.499 persons as the lowest³] [that after *SAB* rising to turn that] [*they* estimated but⁴] [already the estimate misses a point⁵] [prediction became.]

15 In contrast, a two-sentence translation of the Japanese sentence produced by a professional translator ("text(3)") reads as follows:

20 (3) [In its future population estimates¹] [made public last year,²] [the Ministry of Health and Welfare predicted that the SAB would drop to a new low of 1.499 in the future,³] [but would make a comeback after that,⁴] [increasing once again.⁵]
25 [However, it looks as if that prediction will be quickly shattered.⁶]

The labeled spans of text represent elementary discourse units (*edus*), i.e., minimal text spans that have an unambiguous discourse function (Mann et al., "Rhetorical structure theory: Toward a functional theory of text organization," *Text*, 8(3):243-281, 1988). If the text fragments are analyzed closely,

one notices that in translating text (1), a professional translator chose to realize the information in Japanese unit 2 first (unit 2 in text (1) corresponds roughly to unit 1 in text (3)); to realize then some of the information in Japanese unit 1 5 (part of unit 1 in text (1) corresponds to unit 2 in text (3)); to refuse then information given in units 1, 3 and 5 in text (1) and realize it in English as unit 3; and so on. Also, the translator chose to re-package the information in the original Japanese sentence into two English sentences.

10 At the elementary unit level, the correspondence between the Japanese sentence in text (1) and its English translation in text (3) can be represented as in mappings (4), below, where $j \subset e$ denotes the fact that the semantic content of unit j is realized fully in unit e ; $j \supset e$ denotes the fact that the semantic 15 content of unit e is realized fully in unit j ; $j = e$ denotes the fact that units j and e are semantically equivalent; and $j \approx e$ denotes the fact that there is a semantic overlap between units j and e , but neither proper inclusion nor proper equivalence.

20 (4)
$$\begin{aligned} j_1 &\supset e_2; j_1 \approx e_3; \\ j_2 &= e_1; \\ j_3 &\subset e_1; \\ j_4 &\approx e_4; j_4 \approx e_5; \\ j_5 &\approx e_3; \\ j_6 &\subset e_6; \\ j_7 &\subset e_6 \end{aligned}$$

Hence, the mappings in (4) provide an explicit representation of the way information is re-ordered and re-packaged when translated from Japanese into English. However, when translating text, it is also the case that the rhetorical rendering changes. What is realized in Japanese using a CONTRAST relation can be realized in English using, for example, a COMPARISON OR A CONCESSION relation.

Figures 8 and 9 present in the style of Mann, *supra*, the discourse structures of text fragments (1) and (3), above. Each discourse structure is a tree whose leaves correspond to contiguous text spans. Each node is characterized by a *status* (NUCLEUS or SATELLITE) and a *rhetorical relation*, which is a relation that holds between two non-overlapping text spans. The distinction between nuclei and satellites comes from the empirical observation that the nucleus expresses what is more essential to the writer's intention than the satellite; and that the nucleus of a rhetorical relation is comprehensible independent of the satellite, but not vice versa. When spans are equally important, the relation is multinuclear; for example, the CONTRAST relation that holds between unit [3] and span [4,5] in the rhetorical structure of the English text in Figures 8 and 9 is multinuclear. Rhetorical relations that end in the suffix "-e" denote relations that correspond to embedded syntactic constituents. For example, the ELABORATION-OBJECT-ATTRIBUTE-E

relation that holds between units 2 and 1 in the English discourse structure corresponds to a restrictive relative.

If one knows the mappings at the *edu* level, one can determine the mappings at the *span* (discourse constituent) level 5 as well. For example, using the elementary mappings in (4), one can determine that Japanese span [1,2] corresponds to the English span [1,2], Japanese unit [4] to English span [4,5], Japanese span [6,7] to English unit [6], Japanese span [1,5] to English span [1,5], and so on. As Figures 8 and 9 show, the CONCESSION relation that holds between spans [1,5] and [6,7] in the Japanese tree corresponds to a similar relation that holds between span 10 [1,5], and unit [6] in the English tree (modulo the fact that, in Japanese, the relation holds between sentence fragments, while in English it holds between full sentences). However, the TEMPORAL- 15 AFTER relation that holds between units [3] and [4] in the Japanese tree is realized as a CONTRAST relation between unit [3] and span [4,5] in the English tree. And because Japanese units {6} and {7} are fused into unit [6] in English, the relation ELABORATION-OBJECT-ATTRIBUTE-E is no longer made explicit in the 20 English text.

Some of the differences between the two discourse trees in Figures 8 and 9 have been traditionally addressed in MT systems at the syntactic level. For example, the re-ordering of units 1 and 2 can be dealt with using only syntactic models. However, as

discussed below, there are significant differences between Japanese and English with respect to the way information is packaged and organized rhetorically not only at the sentence level, but also, at the paragraph and text levels. More 5 specifically, as humans translate Japanese into English, they re-order the clauses, sentences, and paragraphs of Japanese texts, they re-package the information into clauses, sentences and paragraphs that are not a one-to-one mapping of the original Japanese units, and they rhetorically re-organize the structure of the translated text so as to reflect rhetorical constraints specific to English. If a translation system is to produce text that is not only grammatical but also coherent, it will have to ensure that the discourse structure of the target text reflects the natural renderings of the target language, and not that of the source language.

10 In the Experiment section below, it is empirically shown that there are significant differences between the rhetorical structure of Japanese texts and their corresponding English translations. These differences further demonstrate the need and 20 desirability of developing computational models for discourse structure rewriting.

Experiment

In order to assess the role of discourse structure in MT, a corpus of discourse trees was manually built for 40 Japanese texts and their corresponding translations. The texts were 5 selected randomly from the ARPA corpus (White et al., "Evaluation in the ARPA machine-translation program: 1993 methodology," In *Proceedings of the ARPA Human Language Technology Workshop*, pages 135-140, Washington, D.C., 1994). On average, each text had about 460 words. The Japanese texts had a 10 total of 335 paragraphs and 773 sentences. The English texts had a total of 337 paragraphs and 827 sentences.

A discourse annotation protocol was developed for Japanese and English along the lines followed by Marcu et al., "Experiments in constructing a corpus of discourse trees," In 15 *Proc. Of the ACL'99 Workshop on Standards and Tools for Discourse Tagging*, pages 48-57, Maryland (1999). Marcu's discourse annotation tool (1999) was used in order to manually construct the discourse structure of all Japanese and English texts in the corpus. Ten percent of the Japanese and English texts were 20 rhetorically labeled by two annotators. The tool and the annotation protocol are available at

<http://www.isi.edu/~marcu/software/>

The annotation procedure yielded over the entire corpus 2641
Japanese edus and 2363 English edus.

Corpus	k_u (#)	k_s (#)	k_n (#)	k_r (#)
Japanese	0.856 (80)	0.785 (3377)	0.724 (3377)	0.650 (3377)
English	0.925 (60)	0.866 (1826)	0.839 (1826)	0.748 (1826)

5

Table 1: Tagging reliability

The reliability of the annotation was computer using Marcu
10 et al. (1999)'s method for computing kappa statistics (Siegel et
al., *Non-parametric Statistics for the Behavioral Sciences*,
McGraw-Hill, Second edition, 1988) over hierarchical structures.
Table 1 above displays average kappa statistics that reflect the
reliability of the annotation of elementary discourse units, k_u ,
15 hierarchical discourse spans, k_s , hierarchical nuclearity
assignments, k_n , and hierarchical rhetorical relation
assignments, k_r . Kappa figures higher than 0.8 correspond to
good agreement; kappa figures higher than 0.6 correspond to
acceptable agreement. All kappa statistics were statistically
20 significant at levels higher than $\alpha = 0.01$. In addition to the
kappa statistics, table 1 also displays in parentheses the
average number of data points per document, over which the kappa
statistics were computed.

For each pair of Japanese-English discourse structures, an alignment file was also built manually, which specified in the notation discussed on page 1 the correspondence between the *edus* of the Japanese text and the *edus* of the English translation.

5 The similarity between English and Japanese discourse trees was computed using labeled recall and precision figures that reflected the resemblance of the Japanese and English discourse structures with respect to their assignment of *edu* boundaries, hierarchical spans, nuclearity, and rhetorical relations.

10 Because the trees compared differ from one language to the other in the number of elementary units, the order of these units, and the way the units are grouped recursively into discourse spans, two types of recall and precision figures were computed. In computing *Position-Dependent* (P-D) recall and 15 precision figures, a Japanese span was considered to match an English span when the Japanese span contained all the Japanese *edus* that corresponded to the *edus* in the English span, and when the Japanese and English spans appeared in the same position with respect to the overall structure. For example, the English tree 20 in Figures 8 and 9 are characterized by 10 subsentential spans: [1], [2], [3], [4], [5], [6], [1,2], [4,5], [3,5], and [1,5]. (Span [1,6] subsumes 2 sentences, so it is not sub-sentential.) The Japanese discourse tree has only 4 spans that could be matched in the same positions with English spans, namely spans

[1,2], [4], [5], and [1,5]. Hence the similarity between the Japanese tree and the English tree with respect to their discourse below the sentence level has a recall of 4/10 and a precision of 4/11 (in Figures 8 and 9, there are 11 sub-sentential Japanese spans).

In computing Position-Independent (P-I) recall and precision figures, even when a Japanese span "floated" during the translation to a position in the English tree, the P-I recall and precision figures were not affected. The Position-Independent figures reflect the intuition that if two trees t_1 and t_2 both have a subtree t , t_1 and t_2 are more similar than if they were if they didn't share a tree. At the sentence level, it is assumed that if, for example, the syntactic structure of a relative clause is translated appropriately (even though it is not appropriately attached), this is better than translating wrongly that clause. The Position-Independent figures offer a more optimistic metric for comparing discourse trees. They span a wider range of values than the Position-Dependent figures, which enables a better characterization of the differences between Japanese and English discourse structures. When one takes an optimistic stance, for the spans at the sub-sentential level in the trees in Table 1 the recall is 6/10 and the precision is 6/11 because in addition to spans [1,2], [4], [5], and [1,5], one can

also match Japanese span [1] to English span [2] and Japanese span [2] to Japanese span [1].

In order to provide a better estimate of how close two discourse trees were, Position-Dependent and -Independent recall

5 and precision figures were computed for the sentential level (where units are given by edus and spans are given by sets of edus or single sentences); paragraph level (where units are given by sentences and spans are given by sets of sentences or single paragraphs); and text level (where units are given by paragraphs and spans are given by sets of paragraphs). These figures offer 10 a detailed picture of how discourse structures and relations are mapped from one language to the other across all discourse levels, from sentence to text. The differences at the sentence level can be explained by differences between the syntactic 15 structures of Japanese and English. The differences at the paragraph and text levels have a purely rhetorical explanation.

When the recall and precision figures were computed with respect to the nuclearity and relation assignments, the statuses and the rhetorical relations that labeled each pair of spans were 20 also factored in.

Level	Units		Spans		Status/Nuclearity		Relations	
	P-D R	P-D P	P-D R	P-D P	P-D R	P-D P	P-D R	P-D P
Sentence	29.1	25.0	27.2	22.7	21.3	17.7	14.9	12.4
Paragraph	53.9	53.4	46.8	47.3	38.6	39.0	31.9	32.3
Text	41.3	42.6	31.5	32.6	28.8	29.9	26.1	27.1
Weighted Average	36.0	32.5	31.8	28.4	26.0	23.1	20.1	17.9
All	8.2	7.4	5.9	5.3	4.4	3.9	3.3	3.0

	P-I R	P-IP	P-IP	P-I R	P-IP	P-I R	P-IP	
Sentence	71.0	61.0	56.0	46.6	44.3	36.9	30.5	25.4
Paragraph	62.1	61.6	53.2	53.8	43.3	43.8	35.1	35.5
Text	74.1	76.5	54.4	56.5	48.5	50.4	41.1	42.7
Weighted Average	69.6	63.0	55.2	49.2	44.8	39.9	33.1	29.5
All	74.5	66.8	50.6	45.8	39.4	35.7	26.8	24.3

Table 2: Similarity of Japanese and English discourse structures

Table 2 above summarizes the results (P-D and P-I (R) recall and (P)recision figures) for each level (Sentence, Paragraph, and Text). The numbers in the "Weighted Average" line report averages of the Sentence-, Paragraph-, and Text-specific figures, weighted according to the number of units at each level. The numbers in the "All" line reflect recall and precision figures computed across the entire trees, with no attention paid to sentence and paragraph boundaries.

Given the significantly different syntactic structures of Japanese and English the recall and precision results were low, reflecting the similarity between discourse trees built below the sentence level. However, as Table 2 shows, these are significant differences between discourse trees at the paragraph and text levels as well. For example, the Position-Independent figures show that only about 62% of the sentences and only about 53% of

the hierarchical spans built across sentences could be matched between the two corpora. When one looks at the status and rhetorical relations associated with the spans built across sentences at the paragraph level, the P-I recall and precision 5 figures drop to about 43% and 35% respectively.

The differences in recall and precision are explained both by differences in the way information is packaged into paragraphs in the two languages and the way it is structured rhetorically both within and about the paragraph level.

These results strongly suggest that if one attempts to translate Japanese into English on a sentence-by-sentence basis, it is likely that the resulting text will be unnatural from a discourse perspective. For example, if some information rendered using a CONTRAST relation in Japanese is rendered using an 15 ELABORATION relation in English, it would be inappropriate to use a discourse marker like "but" in the English translation, although that would be consistent with the Japanese discourse structure.

An inspection of the rhetorical mappings between Japanese 20 and English revealed that some Japanese rhetorical renderings are consistently mapped into one or a few preferred renderings in English. For example, 34 of 115 CONTRAST relations in the Japanese texts are mapped into CONTRAST relations in English; 27 become nuclei of relations such as ANTITHESIS and CONCESSION, 14

are translated as COMPARISON relations, 6 as satellites of CONCESSION relations, 5 as LIST relations, etc.

The discourse-based transfer model

5 Figure 10 is a block diagram of a tree rewriter in the process of being trained. As shown therein, the tree rewriter 700 takes as input two different types of trees, for example, a tree of type A and another tree of type B, and automatically learns how to rewrite type A trees into type B trees. The tree rewriter 700 produces as output a transfer function, $H(A \rightarrow B)$, 10 for rewriting trees of type A into trees of type B. Accordingly, assuming type A corresponds to trees in Japanese and type B corresponds to trees in English, $H(A \rightarrow B)$ enables a user (e.g., either a human end-user or a software application invoking the 15 tree rewriter) to automatically convert any tree structure in English into a counterpart tree structure in Japanese.

The tree rewriter works on syntactic trees, rhetorical trees, and virtually any other type of tree structure used in computational linguistics. The tree rewriter has applications 20 not only in machine translation, but also in summarization, discourse parsing, syntactic parsing, information retrieval, automatic test scoring and other applications that generate or use trees. In machine translation, for example, the tree

rewriter can be used to rewrite a syntactic / rhetorical tree specific to one language into a syntactic / rhetorical tree for another language. In summarization, the tree rewriter can be used to rewrite the discourse / syntactic structure of a long 5 text or sentence into the discourse / syntactic structure of a short text or sentence.

This high degree of general applicability is shown in Fig. 11, in which the tree rewriter 801, after it has been trained to learn the transfer function $H(\text{Tree} \rightarrow \text{Tree}')$, can accept a tree 10 as input from any application 800 that generates trees as output. At the output side, the tree rewriter's output (tree' - a rewritten version of the input tree) can be used as input to any application that uses trees as input.

Figure 12 shows a block diagram of a specific application 15 for the tree rewriter as a component of a larger system - namely, a discourse-based machine translation system. Unlike conventional MT systems which in effect adopt a "tiled" approach to translation, for example, by translating individual sentences of a larger work (e.g., a treatise) separately, the 20 discourse-based MT system of Figure 12 translates the entire text as whole, potentially resulting in a translation that has a different number and/or arrangement of sentences than the original, but which better captures the underlying discourse, or rhetoric, of the original text.

As shown in Figure 12, the discourse-based MT system 910 receives as input a source language text 900 and produces as output a target language text 908, which is a discourse-based translation of the source language text 900. The system 910 5 includes three basic components - a discourse parser 902, a discourse-structure transfer module 904 (i.e., a specific instance of a tree rewriter that has been trained to rewrite trees using the transfer function $H(\text{Tree} \rightarrow \text{Tree}')$) and a target language tree-text mapper 906.

10 The discourse parser 902 initially derives the discourse structure of the source language text and produces as output a corresponding discourse tree. Details of a discourse parser that can be used as discourse parser 902 are set forth in Daniel Marcu, "A Decision-Based Approach to Rhetorical Parsing," 15 Proceedings of ACL'99 (1999), which is incorporated herein.

10 The target language tree-text mapper 906 is a statistical module that maps the input text into the target language using translation and language models that incorporate discourse-specific features, which are extracted from the outputs of the discourse parser 209 and the discourse-structure transfer module 904. Details of a suitable mapper 906 are set forth in Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, Kenji Yamada, "Fast Decoding and Optimal Decoding for Machine Translation," Proceedings of the 39th Annual Meeting of the Association for

Computational Linguistics, July 6-11, 2001, which is incorporated herein.

As noted above, the discourse-structure transfer module 904 is a specific instance of a tree rewriter that has been trained 5 to rewrite trees from the desired input type into the desired output type. More specifically, the discourse-structure transfer module 904 rewrites the discourse structure of the input text so as to reflect a discourse rendering that is natural to the target text.

10 Fig. 13 is a flow chart illustrating a process 1300 that can be used to train a tree rewriter to automatically learn the transfer function between two different types of tree structures, e.g., a tree of type A and a tree of type B.

15 As shown in Fig. 13, the first step 1301 is to generate a training set of input-output pairs of trees $[T_s, T_t]$ and a mapping C between leaves each input-output tree pair. The input tree of the pair is of the type from which conversion is desired or, in other words, the source tree type T_s . The output tree of the pair is of the type to which conversion is desired or, in 20 other words, the target tree type T_t .

The mapping C between leaves of an input tree and its paired output tree defines a correspondence between a source text segment and its counterpart target language translation. These mappings either can be generated manually as described below, or

automatically, as described in Kevin Knight and Daniel Marcu,
"Statistics-Based Summarization - Step One: Sentence
Compression," The 17th National Conference on Artificial
Intelligence (AAAI--2000), pp. 703-710, which is incorporated by
5 reference.

Available types of mappings between leaves of a Japanese-
English input-output pair that can be used are shown in equation
10 (4) above, wherein j refers to a Japanese text segment and e
refers to an English translation of that text segment. Note that
the mappings represented by equation (4) are not limited to one-
15 to-one mappings but rather can be any arbitrary mapping - that
is, not only one-to-one, but also one-to-many, many-to-one and
many-to-many. This flexibility in the mappings dramatically
enhances the richness with which relationships between the input
and output trees are defined, and further enhances the
flexibility of the transfer function $H[]$ that is automatically
20 learned.

After the training set (input-output tree pairs and mappings
therebetween) has been generated, the training process next
25 determines in step 1303 the grouping and order of operations that
generate a given output tree starting with its paired input tree.
This step is performed based on the following seven basic
operations, referred to collectively as the "extended shift-
reduce" operations - shift, reduce, break, create-next, fuse,

swap, and assignType - which are described in detail below under the section heading "The discourse-based transfer model."

These seven operations are sufficient to rewrite any given input tree into its paired output tree.

5 The output of step 1303 is a set of learning cases - one learning case for each input-output tree pair in the training set. In essence, each learning case is an ordered set of extended shift-reduce operations that when applied to an input tree will generate the paired output tree.

10 Next, in step 1305, the tree rewriter training process 1300 associates features (e.g., operational and discourse features, correspondence-based features, and lexical features) with the learning cases to reflect the context in which the operations are to be performed. Details of step 1305 are discussed below under 15 the heading "Learning the parameters of the discourse-transfer model."

Next, in step 1307, the tree rewriter training process 1300 applies a learning algorithm, for example, the C4.5 algorithm as described in J. Ross Quinlan, "C4.5: Programs for Machine 20 Learning," Morgan Kaufmann Publishers (1993), to learn a set of decision rules from the learning cases. Details of step 1307 are discussed below under the heading "Learning the parameters of the discourse-transfer model." This set of decision rules collectively constitute the transfer function, $H(T_s \rightarrow T_t)$, for

rewriting any tree of type T_s into a tree of type T_t . This transfer function can then be used by users, applications or other automated processes for rewriting previously unseen trees of type T_s into trees of type T_t .

5 A more detailed discussion of training a tree rewriter follows.

In order to learn to rewrite discourse structure trees, a related problem, defined below in Definition 3.1, is addressed.

0
010
015
020
025
030
035
040
045
050
055
060
065
070
075
080
085
090
095
100
105
110
115
120
125
130
135
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215
220
225
230
235
240
245
250
255
260
265
270
275
280
285
290
295
300
305
310
315
320
325
330
335
340
345
350
355
360
365
370
375
380
385
390
395
400
405
410
415
420
425
430
435
440
445
450
455
460
465
470
475
480
485
490
495
500
505
510
515
520
525
530
535
540
545
550
555
560
565
570
575
580
585
590
595
600
605
610
615
620
625
630
635
640
645
650
655
660
665
670
675
680
685
690
695
700
705
710
715
720
725
730
735
740
745
750
755
760
765
770
775
780
785
790
795
800
805
810
815
820
825
830
835
840
845
850
855
860
865
870
875
880
885
890
895
900
905
910
915
920
925
930
935
940
945
950
955
960
965
970
975
980
985
990
995
1000
1005
1010
1015
1020
1025
1030
1035
1040
1045
1050
1055
1060
1065
1070
1075
1080
1085
1090
1095
1100
1105
1110
1115
1120
1125
1130
1135
1140
1145
1150
1155
1160
1165
1170
1175
1180
1185
1190
1195
1200
1205
1210
1215
1220
1225
1230
1235
1240
1245
1250
1255
1260
1265
1270
1275
1280
1285
1290
1295
1300
1305
1310
1315
1320
1325
1330
1335
1340
1345
1350
1355
1360
1365
1370
1375
1380
1385
1390
1395
1400
1405
1410
1415
1420
1425
1430
1435
1440
1445
1450
1455
1460
1465
1470
1475
1480
1485
1490
1495
1500
1505
1510
1515
1520
1525
1530
1535
1540
1545
1550
1555
1560
1565
1570
1575
1580
1585
1590
1595
1600
1605
1610
1615
1620
1625
1630
1635
1640
1645
1650
1655
1660
1665
1670
1675
1680
1685
1690
1695
1700
1705
1710
1715
1720
1725
1730
1735
1740
1745
1750
1755
1760
1765
1770
1775
1780
1785
1790
1795
1800
1805
1810
1815
1820
1825
1830
1835
1840
1845
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
1900
1905
1910
1915
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
2035
2040
2045
2050
2055
2060
2065
2070
2075
2080
2085
2090
2095
2100
2105
2110
2115
2120
2125
2130
2135
2140
2145
2150
2155
2160
2165
2170
2175
2180
2185
2190
2195
2200
2205
2210
2215
2220
2225
2230
2235
2240
2245
2250
2255
2260
2265
2270
2275
2280
2285
2290
2295
2300
2305
2310
2315
2320
2325
2330
2335
2340
2345
2350
2355
2360
2365
2370
2375
2380
2385
2390
2395
2400
2405
2410
2415
2420
2425
2430
2435
2440
2445
2450
2455
2460
2465
2470
2475
2480
2485
2490
2495
2500
2505
2510
2515
2520
2525
2530
2535
2540
2545
2550
2555
2560
2565
2570
2575
2580
2585
2590
2595
2600
2605
2610
2615
2620
2625
2630
2635
2640
2645
2650
2655
2660
2665
2670
2675
2680
2685
2690
2695
2700
2705
2710
2715
2720
2725
2730
2735
2740
2745
2750
2755
2760
2765
2770
2775
2780
2785
2790
2795
2800
2805
2810
2815
2820
2825
2830
2835
2840
2845
2850
2855
2860
2865
2870
2875
2880
2885
2890
2895
2900
2905
2910
2915
2920
2925
2930
2935
2940
2945
2950
2955
2960
2965
2970
2975
2980
2985
2990
2995
3000
3005
3010
3015
3020
3025
3030
3035
3040
3045
3050
3055
3060
3065
3070
3075
3080
3085
3090
3095
3100
3105
3110
3115
3120
3125
3130
3135
3140
3145
3150
3155
3160
3165
3170
3175
3180
3185
3190
3195
3200
3205
3210
3215
3220
3225
3230
3235
3240
3245
3250
3255
3260
3265
3270
3275
3280
3285
3290
3295
3300
3305
3310
3315
3320
3325
3330
3335
3340
3345
3350
3355
3360
3365
3370
3375
3380
3385
3390
3395
3400
3405
3410
3415
3420
3425
3430
3435
3440
3445
3450
3455
3460
3465
3470
3475
3480
3485
3490
3495
3500
3505
3510
3515
3520
3525
3530
3535
3540
3545
3550
3555
3560
3565
3570
3575
3580
3585
3590
3595
3600
3605
3610
3615
3620
3625
3630
3635
3640
3645
3650
3655
3660
3665
3670
3675
3680
3685
3690
3695
3700
3705
3710
3715
3720
3725
3730
3735
3740
3745
3750
3755
3760
3765
3770
3775
3780
3785
3790
3795
3800
3805
3810
3815
3820
3825
3830
3835
3840
3845
3850
3855
3860
3865
3870
3875
3880
3885
3890
3895
3900
3905
3910
3915
3920
3925
3930
3935
3940
3945
3950
3955
3960
3965
3970
3975
3980
3985
3990
3995
4000
4005
4010
4015
4020
4025
4030
4035
4040
4045
4050
4055
4060
4065
4070
4075
4080
4085
4090
4095
4100
4105
4110
4115
4120
4125
4130
4135
4140
4145
4150
4155
4160
4165
4170
4175
4180
4185
4190
4195
4200
4205
4210
4215
4220
4225
4230
4235
4240
4245
4250
4255
4260
4265
4270
4275
4280
4285
4290
4295
4300
4305
4310
4315
4320
4325
4330
4335
4340
4345
4350
4355
4360
4365
4370
4375
4380
4385
4390
4395
4400
4405
4410
4415
4420
4425
4430
4435
4440
4445
4450
4455
4460
4465
4470
4475
4480
4485
4490
4495
4500
4505
4510
4515
4520
4525
4530
4535
4540
4545
4550
4555
4560
4565
4570
4575
4580
4585
4590
4595
4600
4605
4610
4615
4620
4625
4630
4635
4640
4645
4650
4655
4660
4665
4670
4675
4680
4685
4690
4695
4700
4705
4710
4715
4720
4725
4730
4735
4740
4745
4750
4755
4760
4765
4770
4775
4780
4785
4790
4795
4800
4805
4810
4815
4820
4825
4830
4835
4840
4845
4850
4855
4860
4865
4870
4875
4880
4885
4890
4895
4900
4905
4910
4915
4920
4925
4930
4935
4940
4945
4950
4955
4960
4965
4970
4975
4980
4985
4990
4995
5000
5005
5010
5015
5020
5025
5030
5035
5040
5045
5050
5055
5060
5065
5070
5075
5080
5085
5090
5095
5100
5105
5110
5115
5120
5125
5130
5135
5140
5145
5150
5155
5160
5165
5170
5175
5180
5185
5190
5195
5200
5205
5210
5215
5220
5225
5230
5235
5240
5245
5250
5255
5260
5265
5270
5275
5280
5285
5290
5295
5300
5305
5310
5315
5320
5325
5330
5335
5340
5345
5350
5355
5360
5365
5370
5375
5380
5385
5390
5395
5400
5405
5410
5415
5420
5425
5430
5435
5440
5445
5450
5455
5460
5465
5470
5475
5480
5485
5490
5495
5500
5505
5510
5515
5520
5525
5530
5535
5540
5545
5550
5555
5560
5565
5570
5575
5580
5585
5590
5595
5600
5605
5610
5615
5620
5625
5630
5635
5640
5645
5650
5655
5660
5665
5670
5675
5680
5685
5690
5695
5700
5705
5710
5715
5720
5725
5730
5735
5740
5745
5750
5755
5760
5765
5770
5775
5780
5785
5790
5795
5800
5805
5810
5815
5820
5825
5830
5835
5840
5845
5850
5855
5860
5865
5870
5875
5880
5885
5890
5895
5900
5905
5910
5915
5920
5925
5930
5935
5940
5945
5950
5955
5960
5965
5970
5975
5980
5985
5990
5995
6000
6005
6010
6015
6020
6025
6030
6035
6040
6045
6050
6055
6060
6065
6070
6075
6080
6085
6090
6095
6100
6105
6110
6115
6120
6125
6130
6135
6140
6145
6150
6155
6160
6165
6170
6175
6180
6185
6190
6195
6200
6205
6210
6215
6220
6225
6230
6235
6240
6245
6250
6255
6260
6265
6270
6275
6280
6285
6290
6295
6300
6305
6310
6315
6320
6325
6330
6335
6340
6345
6350
6355
6360
6365
6370
6375
6380
6385
6390
6395
6400
6405
6410
6415
6420
6425
6430
6435
6440
6445
6450
6455
6460
6465
6470
6475
6480
6485
6490
6495
6500
6505
6510
6515
6520
6525
6530
6535
6540
6545
6550
6555
6560
6565
6570
6575
6580
6585
6590
6595
6600
6605
6610
6615
6620
6625
6630
6635
6640
6645
6650
6655
6660
6665
6670
6675
6680
6685
6690
6695
6700
6705
6710
6715
6720
6725
6730
6735
6740
6745
6750
6755
6760
6765
6770
6775
6780
6785
6790
6795
6800
6805
6810
6815
6820
6825
6830
6835
6840
6845
6850
6855
6860
6865
6870
6875
6880
6885
6890
6895
6900
6905
6910
6915
6920
6925
6930
6935
6940
6945
6950
6955
6960
6965
6970
6975
6980
6985
6990
6995
7000
7005
7010
7015
7020
7025
7030
7035
7040
7045
7050
7055
7060
7065
7070
7075
7080
7085
7090
7095
7100
7105
7110
7115
7120
7125
7130
7135
7140
7145
7150
7155
7160
7165
7170
7175
7180
7185
7190
7195
7200
7205
7210
7215
7220
7225
7230
7235
7240
7245
7250
7255
7260
7265
7270
7275
7280
7285
7290
7295
7300
7305
7310
7315
7320
7325
7330
7335
7340
7345
7350
7355
7360
7365
7370
7375
7380
7385
7390
7395
7400
7405
7410
7415
7420
7425
7430
7435
7440
7445
7450
7455
7460
7465
7470
7475
7480
7485
7490
7495
7500
7505
7510
7515
7520
7525
7530
7535
7540
7545
7550
7555
7560
7565
7570
7575
7580
7585
7590
7595
7600
7605
7610
7615
7620
7625
7630
7635
7640
7645
7650
7655
7660
7665
7670
7675
7680
7685
7690
7695
7700
7705
7710
7715
7720
7725
7730
7735
7740
7745
7750
7755
7760
7765
7770
7775
7780
7785
7790
7795
7800
7805
7810
7815
7820
7825
7830
7835
7840
7845
7850
7855
7860
7865
7870
7875
7880
7885
7890
7895
7900
7905
7910
7915
7920
7925
7930
7935
7940
7945
7950
7955
7960
7965
7970
7975
7980
7985
7990
7995
8000
8005
8010
8015
8020
8025
8030
8035
8040
8045
8050
8055
8060
8065
8070
8075
8080
8085
8090
8095
8100
8105
8110
8115
8120
8125
8130
8135
8140
8145
8150
8155
8160
8165
8170
8175
8180
8185
8190
8195
8200
8205
8210
8215
8220
8225
8230
8235
8240
8245
8250
8255
8260
8265
8270
8275
8280
8285
8290
8295
8300
8305
8310
8315
8320
8325
8330
8335
8340
8345
8350
8355
8360
8365
8370
8375
8380
8385
8390
8395
8400
8405
8410
8415
8420
8425
8430
8435
8440
8445
8450
8455
8460
8465
8470
8475
8480
8485
8490
8495
8500
8505
8510
8515
8520
8525
8530
8535
8540
8545
8550
8555
8560
8565
8570
8575
8580
8585
8590
8595
8600
8605
8610
8615
8620
8625
8630
8635
8640
8645
8650
8655
8660
8665
8670
8675
8680
8685
8690
8695
8700
8705
8710
8715
8720
8725
8730
8735
8740
8745
8750
8755
8760
8765
8770
8775
8780
8785
8790
8795
8800
8805
8810
8815
8820
8825
8830
8835
8840
8845
8850
8855
8860
8865
8870
8875
8880
8885
8890
8895
8900
8905
8910
8915
8920
8925
8930
8935
8940
8945
8950
8955
8960
8965
8970
8975
8980
8985
8990
8995
9000
9005
9010
9015
9020
9025
9030
9035
9040
9045
9050
9055
9060
9065
9070
9075
9080
9085
9090
9095
9100
9105
9110
9115
9120
9125
9130
9135
9140
9145
9150
9155
9160
9165
9170
9175
9180
9185
9190
9195
9200
9205
9210
9215
9220
9225
9230
9235
9240
9245
9250
9255
9260
9265
9270
9275
9280
9285
9290
9295
9300
9305
9310
9315
9320
9325
9330
9335
9340
9345
9350
9355
9360
9365
9370
9375
9380
9385
9390
9395
9400
9405
9410
9415
9420
9425
9430
9435
9440
9445
9450
9455
9460
9465
9470
9475
9480
9485
9490
9495
9500
9505
9510
9515
9520
9525
9530
9535
9540
9545
9550
9555
9560
9565
9570
9575
9580
9585
9590
9595
9600
9605
9610
9615
9620
9625
9630
9635
9640
9645
9650
9655
9660
9665
9670
9675
9680
9685
9690
9695
9700
9705
9710
9715
9720
9725
9730
9735
9740
9745
9750
9755
9760
9765
9770
9775
9780
9785
9790
9795
9800
9805
9810
9815
9820
9825
9830
9835
9840
9845
9850
9855
9860
9865
9870
9875
9880
9885
9890
9895
9900
9905
9910
9915
9920
9925
9930
9935
9940
9945
9950
9955
9960
9965
9970
9975
9980
9985
9990
9995
10000
10005
10010
10015
10020
10025
10030
10035
10040
10045
10050
10055
10060
10065
10070
10075
10080
10085
10090
10095
10100
10105
10110
10115
10120
10125
10130
10135
10140
10145
10150
10155
10160
10165
10170
10175
10180
10185
10190
10195
10200
10205
10210<br

Spain (1997), and Marcu, "A decision-based approach to rhetorical parsing," In Proc. Of ACL'99, pages 365-372, Maryland (1999). In this extended paradigm, the transfer process starts with an empty Stack and an Input List that contains a sequence of 5 elementary discourse trees $edts$, one edt for each edu in the tree T_s given as input. The status and rhetorical relation associated with each edt is undefined. At each step, the transfer module applies an operation that is aimed at building from the units in 10 T_s the discourse tree T_t . In the context of the discourse-transfer module, 7 types of operations are implemented:

- SHIFT operations transfer the first edt from the input list into the stack;
- REDUCE operations pop the two discourse trees located at the top of the stack; combine them into a new tree updating the statuses and rhetorical relation names of the trees involved in the operation; and push the new tree on the top of the stack. These operations are used to build the structure of the discourse tree in the target language.
- 20 • BREAK operations are used in order to break the edt at the beginning of the input list into a predetermined number of units. These operations are used to ensure that the resulting tree has the same number of $edts$ as T_t . For

example, a BREAK operation is used whenever a Japanese *edu* is mapped into multiple English units.

- CREATE-NEXT operations are used, for example, in order to create English (target language) discourse constituents that have no correspondent in the Japanese (source language) tree.
- FUSE operations are used in order to fuse the *edt* at the top of the stack into the tree that immediately precedes it. These operations are used, for example, whenever multiple Japanese *edus* are mapped into one English *edu*.
- SWAP operations swap the *edt* at the beginning of the input list with an *edt* found one or more positions to the right. These operations are used for re-ordering discourse constituents.
- ASSIGNTYPE operations assign one or more of the following types to the tree at the top of the stack: Unit, MultiUnit, Sentence, Paragraph, MultiParagraph, and Text. These operations are used in order to ensure sentence and paragraph boundaries that are specific to the target language.

For example, the first sentence of the English tree in Figure 9 can be obtained from the original Japanese sequence by

following the sequence of actions (5), below, whose effects are shown in Figure 14. For the purpose of compactness, Figure 14 does not illustrate the effect of ASSIGNTYPE actions. For the same purpose, some lines correspond to more than one action.

5 (5) BREAK 2; SWAP 2; SHIFT; ASSIGNTYPE UNIT;
SHIFT; REDUCE-NS-ELABORATION-OBJECT-
ATTRIBUTE-E; ASSIGNTYPE MULTIUNIT; SHIFT;
ASSIGNTYPE UNIT; SHIFT; ASSIGNTYPE UNIT;
FUSE; ASSIGNTYPE UNIT; SWAP2; SHIFT;
ASSIGNTYPE UNIT; FUSE; BREAK2; SHIFT;
ASSIGNTYPE UNIT; SHIFT; ASSIGNTYPE UNIT;
REDUCE-NS-ELABORATION-ADDITIONAL; ASSIGNTYPE
MULTIUNIT; REDUCE-NS-CONTRAST; ASSIGNTYPE
MULTIUNIT; REDUCE-SN-BACKGROUND; ASSIGNTYPE
SENTENCE.

10
15
20
25
For the corpus used, in order to enable a discourse-based transfer module to derive any English discourse tree starting from any Japanese discourse tree, it is sufficient to implement:

- one SHIFT operation;
- $3 \times 2 \times 85$ REDUCE operations; (For each of the three possible pairs of nuclearity assignments NUCLEUS-SATELLITE (NS), SATELLITE-NUCLEUS (SN), AND NUCLEUS-NUCLEUS (NN), there are two possible ways to reduce two adjacent trees (one results in a binary tree, the other in a non-binary tree (Marcu, "A decision-based approach to rhetorical parsing," In Proc. Of ACL'99, pages 365-372, Maryland (1999)), and 85 relation names.);

- three types of BREAK operations; (In the corpus used, a Japanese unit is broken into two, three, or at most four units.);
- one type of CREATE-NEXT operation;
- one type of FUSE operation;
- eleven types of SWAP operations; (In the corpus, Japanese units are at most 11 positions away from their location in an English-specific rendering.)
- seven types of ASSIGNTYPE operations: Unit, MultiUnit, Sentence, MultiSentence, paragraph, MultiParagraph, and Text.

These actions are sufficient for rewriting any tree T_s into any tree T_t , where T_t may have a different number of edus, where the edus of T_t may have a different ordering than the edus of T_s , and where the hierarchical structures of the two trees may be different as well.

Learning the parameters of the discourse-transfer model

Each configuration of the transfer model is associated with a learning case. The cases were generated by a program that automatically derived the sequence of actions that mapped the Japanese trees in the corpus into the sibling English trees, using the correspondences at the elementary unit level that were

constructed manually. Overall, the 40 pairs of Japanese and English discourse trees yielded 14108 cases.

To each learning example, a set of features from the following classes was associated:

5 **Operational and discourse features** reflect the number of trees in the stack, the input list, and the types of the last five operations. They encode information pertaining to the types of the partial trees built up to a certain time and the rhetorical relations that hold between these trees.

10 **Correspondence-based features** reflect the nuclearity, rhetorical relations, and types of the Japanese trees that correspond to the English-like partial trees derived up to a given time.

15 **Lexical features** specify whether the Japanese spans that correspond to the structures derived up to a given time use potential discourse markers, such as *dakara* (*because*) and *no ni* (*although*).

20 The discourse transfer module uses the C4.5 program (Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers (1993)) in order to learn decision trees and rules that specify how Japanese discourse trees should be mapped into English-like trees. A ten-fold cross-validation evaluation of the classifier yielded an accuracy of 70.2% (± 0.21).

In order to better understand the strengths and weaknesses of the classifier, the problem was broken into smaller components. Hence, instead of learning all actions at once, it was determined first whether the rewriting procedure should

5 choose a SHIFT, REDUCE, BREAK, FUSE, SWAP, or ASSIGNTYPE operation (the "Main Action Type" classifier in table 3), and only then to refine this decision by determining what type of reduce operation to perform how many units to break a Japanese units into, how big the distance to the SWAP-ed unit should be, and what type of ASSIGNTYPE operation one should perform. Table 10 3 below shows the sizes of each data set and the performance of each of these classifiers, as determined using a ten-fold cross-validation procedure. For the purpose of comparison, each classifier is paired with a majority baseline.

15

Classifier	# Cases	Accuracy (10-fold cross validation)	Majority baseline accuracy
General (Learns all classes at once)	14108	70.20% (± 0.21)	22.05% (on ASSIGNTYPE UNIT)
Main Action Type	14108	82.53% (± 0.25)	45.47% (on ASSIGNTYPE)
AssignType	6416	90.46% (± 0.39)	57.30% (on ASSIGNTYPE Unit)
Break	394	82.91% (± 1.40)	82.91% (on BREAK 2)
Nuclearity-Reduce	2388	67.43% (± 1.03)	50.92% (on NS)
Relation-Reduce	2388	48.20% (± 1.01)	17.18% (on ELABORATION- OBJECT-ATTRIBUTE-E)
Swap	842	62.98% (± 1.62)	62.98% (on SWAP1)

Table 3: Performance of the classifiers

The results in Table 3 show that the most difficult subtasks to learn are that of determining the number of units a Japanese unit should be broken into and that of determining the distance to the unit that is to be swapped. The features used are not able to refine the baseline classifiers for these action types. The confusion matrix for the "Main Action Type" classifier (see Table 4) shows that the system has trouble mostly identifying BREAK and CREATE-NEXT actions. The system has difficulty learning what type of nuclearity ordering to prefer (the "Nuclearity-Reduce" classifier) and what relation to choose for the English-like structure (the "Relation-Reduce" classifier).

Action	(a)	(b)	(c)	(d)	(e)	(f)	(g)
ASSIGNTYPE (a)	660						
BREAK (b)		1			2	28	1
CREATE-NEXT (c)					1	8	
FUSE (d)				69	8	3	
REDUCE (e)		4			193	30	3
SHIFT (f)	1	4		15	44	243	25
SWAP (g)		3		4	14	43	25

Table 4: Confusion matrix for the Main Action Type classifier

15

Figure 15 shows a typical learning curve, the one that corresponds to the "Reduce Relation" classifier. The learning curves suggest that more training data may improve performance. 20 However, they also suggest that better features may be needed in order to improve performance significantly.

Table 5 below displays some learned rules. The first rule accounts for rhetorical mappings in which the order of the nucleus and satellite of an ATTRIBUTION relation is changed when translated from Japanese into English. The second rule was 5 learned in order to map EXAMPLE Japanese satellites into EVIDENCE English satellites.

```
10      if rhetRelOfStack-1InJapTree = ATTRIBUTION
            then rhet1OfTopStackInEngTree ATTRIBUTION

            if rhetRelOfTopStackInJapTree = EXAMPLE ∧
                isSentenceTheLastUnitInJapTreeOfTopStack = false
                then rhetRelOfTopStackInEngTree EVIDENCE
15
```

Table 5: Rule examples for the Relation-Reduce classifier

Evaluation of the discourse-based transfer module

By applying the General classifier or the other six classifiers successively, one can map any Japanese discourse tree into a tree whose structure comes closer to the natural rendering of English. To evaluate the discourse-based transfer module, a ten-fold cross-validation experiment was performed. That is, the classifiers were trained on 36 pairs of manually built and aligned discourse structures, and then the learned classifiers were used in order to map 4 unseen Japanese discourse trees into English-like trees. The similarity of the derived trees and the English trees built manually was measured using the metrics

discussed above. This procedure was repeated ten times, each time training and testing on different subsets of tree pairs.

The results reported in Table 2 were as a baseline for the model. The baseline corresponds to applying no knowledge of discourse. Table 6 below displays the absolute improvement (in percentage points) in recall and precision figures obtained when the General classifier was used to map Japanese trees into English-looking trees. The General classifier yielded the best results. The results in Table 6 are averaged over a ten-fold cross-validation experiment.

Level	Units		Spans		Status/Nuclearity		Relations	
	P-D R	P-D P	P-D R	P-D P	P-D R	P-D P	P-D R	P-D P
Sentence	+9.1	+25.5	+2.0	+19.9	+0.4	+13.4	-0.01	+8.4
Paragraph	-14.7	+1.4	-12.5	-1.7	-11.0	-2.4	-9.9	-3.3
Text	-9.6	-13.5	-7.1	-11.1	-6.3	-10.0	-5.2	-8.8
Weighted Average	+1.5	+14.1	-2.1	+9.9	-3.1	+6.4	-3.0	+3.9
All	-1.2	+2.5	-0.1	+2.9	+0.6	+3.5	+0.7	+2.6

	P-I R		P-I P		P-I R		P-I P		P-I R		P-I P	
	P-I R	P-I P										
Sentence	+13.4	+30.4	+3.1	+36.1	-6.3	+18.6	-10.1	+3.9				
Paragraph	-15.6	+0.6	-13.5	-0.8	-11.7	-1.8	-10.3	-2.9				
Text	-15.4	-23.3	-13.0	-20.4	-13.2	-19.5	-11.5	-17.0				
Weighted Average	+3.6	+15.5	-2.7	+17.1	-8.5	+7.3	-10.5	-0.4				
All	+12.7	+29.6	+2.0	+28.8	-5.1	+13.0	-7.9	+2.2				

Table 6: Relative evaluation of the discourse-based transfer module with respect to the figures in Table 2

The results in Table 6 show that the model described here outperforms the baseline with respect to building English-like discourse structures for sentences, but it under-performs the

baseline with respect to building English-like structures at the paragraph and text levels. One potential shortcoming of this model seems to come from its low performance in assigning paragraph boundaries. Because the classifier does not learn correctly which spans to consider paragraphs and which spans not, the recall and precision results at the paragraph and text levels are negatively affected. The poorer results at the paragraph and text levels can be also explained by errors whose effect cumulates during the step-by-step tree-reconstruction procedure; and by the fact that, for these levels, there is less data to learn from.

However, if one ignores the sentence and paragraph boundaries and evaluates the discourse structures overall, one can see that this model outperforms the baseline on all accounts according to the Position-Dependent evaluation; outperforms the baseline with respect to the assignment of elementary units, hierarchical spans, and nuclearity statuses according to the Position-Independent evaluation and under-performs the baseline only slightly with respect to the rhetorical relation assignment according to the Position-Independent evaluation. More sophisticated discourse features, such as those discussed by Maynard, *Principles of Japanese Discourse: A Handbook*, Cambridge Univ. Press (1998), for example, and a tighter integration with

the lexicogrammar of the two languages may yield better cues for learning discourse-based translation models.

Alternative embodiments of the tree rewriter are possible.

For example, probabilities could be incorporated into the tree

5 rewriting procedure. Alternatively, or in addition, multiple trees could be rewritten in parallel and the best one selected in the end. In the current embodiment, the target tree T_t is generated in a sequence of deterministic steps with no recursion or branching. Alternatively, it is possible to associate a probability with each individual step and reconstruct the target tree T_t by exploring multiple alternatives at the same time. The probability of a target tree T_t is given by the product of the probabilities of all steps that led to that tree. In this case, the target tree T_t will be taken to be the resulting tree of 15 maximum probability. An advantage of such an approach is that it enables the learning of probabilistic transfer functions, $H[]$.

Although only a few embodiments have been described in detail above, those having ordinary skill in the art will certainly understand that many modifications are possible in the 20 preferred embodiment without departing from the teachings thereof. All such modifications are encompassed within the following claims.